Self-Replication and Evolution of DNA Crystals

Rebecca Schulman and Erik Winfree

California Institute of Technology, Pasadena, CA 91125, USA
{rebecka,winfree}@caltech.edu

Abstract. Is it possible to create a simple physical system that is capable of replicating itself? Can such a system evolve interesting behaviors, thus allowing it to adapt to a wide range of environments? This paper presents a design for such a replicator constructed exclusively from synthetic DNA. The basis for the replicator is crystal growth: information is stored in the spatial arrangement of monomers and copied from layer to layer by templating. Replication is achieved by fragmentation of crystals, which produces new crystals that carry the same information. Crystal replication avoids intrinsic problems associated with template-directed mechanisms for replication of one-dimensional polymers. A key innovation of our work is that by using programmable DNA tiles as the crystal monomers, we can design crystal growth processes that apply interesting selective pressures to the evolving sequences. While evolution requires that copying occur with high accuracy, we show how to adapt error-correction techniques from algorithmic self-assembly to lower the replication error rate as much as is required.

1 Introduction

It is widely accepted that Darwinian evolution is responsible for the complexity and adaptability seen in modern biology. However, the mechanisms by which evolving organisms adapt to their environment are not well understood. An important roadblock in studying evolution is the dearth of physical systems in which evolution can be studied; a tractable synthetic system for replication and evolution would facilitate the study of how physical selection pressures lead to evolutionary adaptation. A chemical self-replicator might also be used to evolve solutions to problems in chemistry or nanotechnology. If such a system were simple enough, it could also shed light on how self-replication emerged spontaneously at the origin of life.

In 1966, Graham Cairns-Smith proposed a simple mechanism by which polytypic clay crystals could replicate information in the absence of biological enzymes [3, 4]. Polytypic clay crystals are crystals where the orientations of subsequent layers can differ, and therefore a cross-section of the crystal contains an information-bearing sequence. Crystal growth extends the layers and copies the sequence of orientations, which may be considered its genotype. Occasionally, physical forces break a crystal apart. Because crystals replicate their genotype many times during growth, splitting of a crystal can yield multiple pieces, each containing at least one copy of the entire genotype. Cycles of growth and fragmentation cause each sequence to be exponentially amplified.

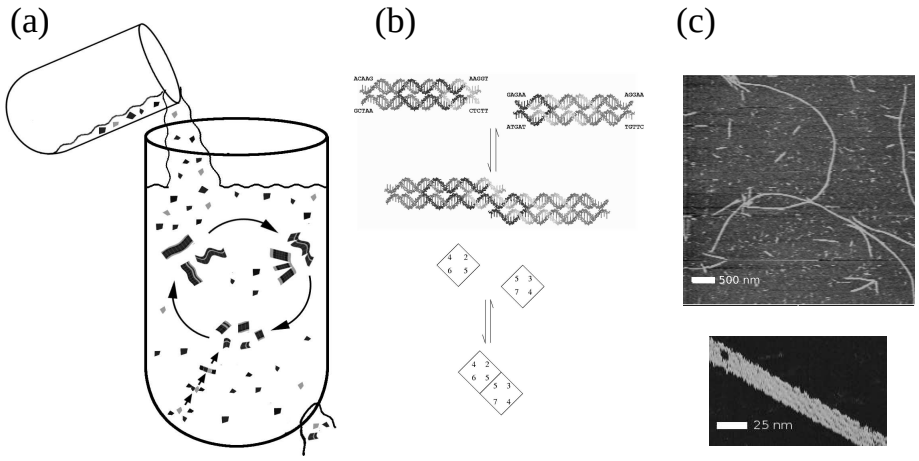

Fig. 1. DNA crystals. (a) The DNA crystal life cycle. The materials required for growth are constantly replenished. Crystals die when they are flushed out of solution in an exit stream. **(b)** Tiles with complementary single-stranded sticky ends can attach by hybridization. For convenience, DNA tiles may be represented as square tiles; tiles with the same side labels correspond to molecules with matching sticky ends. **(c)** Atomic force microscopy image of DNA crystals formed by the molecules shown in Figure 2b. At higher resolutions, individual tiles can be discriminated.

We propose a method of self-replication that works by similar growth and fragmentation of algorithmic DNA crystals. DNA crystals are composed of DNA tile monomers [8]. Different types of DNA tiles can be designed to assemble via programmable rules [18]; a typical DNA crystal is assembled from several tile types. As in Graham-Smith's conception, DNA crystals can contain a sequence that is copied during growth, in this case a linear arrangement of DNA tile types (Figure 1a). Unlike most types of clay crystal growth, DNA crystal growth is tractable in the laboratory and occurs at time scales (hours) that are suitable for experimental investigation.

It is perhaps surprising that DNA crystal replication exhibits many of the phenomena of interest for the study of Darwinian evolution. In Section 2, we describe in more detail how crystal evolution works and introduce the components of DNA crystals and a model of the growth process. The examples in Sections 3 and 4 illustrate how DNA crystals can copy arbitrary amounts of information and how in particular environments, this information affects the replication rate. In Section 5, we describe techniques for increasing the accuracy of replication.

2 Replicating Information with DNA Crystals

DNA crystals consist of DNA tile monomers [8] which can attach to other tiles in a programmable fashion: each of the four sides of the DNA tile has a short single stranded portion which can hybridize with the complementary strand of another tile (Figure 1b). DNA tiles can assemble into 2-dimensional crystals [21] and can be programmed to form other structures, such as thin ribbons (Figure 1c). A wide variety of DNA tile crystals have been synthesized [10, 15, 5].

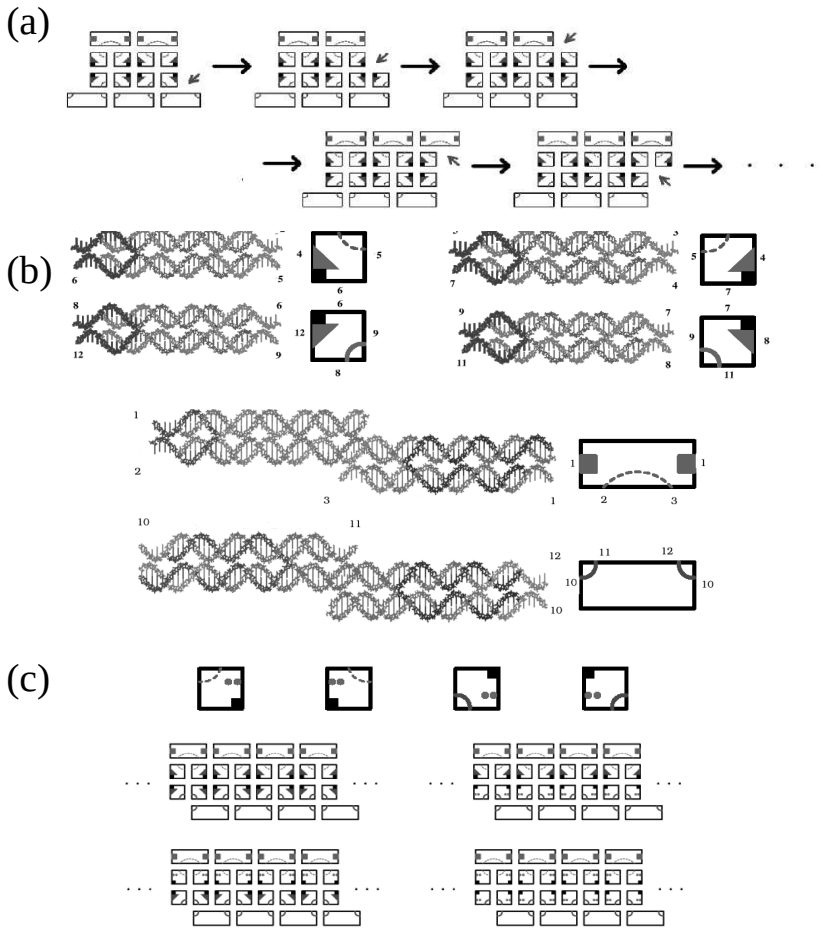

Fig. 2. The zig-zag tile set. (a) A zig-zag assembly. Two alternating tile types in each row enforce the placement of the double tiles on the top and bottom, ensuring that under algorithmic assembly conditions, growth occurs in a zig-zag pattern. Although only growth on the right end of the molecule is shown here, growth occurs simultaneously on both ends of the molecule. At each step, a new tile may be added at the location designated by the small arrow. **(b)** The basic zig-zag tile set consists of six molecules (tile types). Each square and rectangle shown is a logical representation of the molecule shown to its left. By convention, tiles cannot be rotated. The tiles shown here have unique bonds that determine where they fit in the assembly: each label has exactly one match on another tile type. While the logical representations of DNA tiles have the same connectivity as DNA tile molecules, the logical representation of a tile has a different aspect ratio and labels in different orientations than the actual molecules. **(c)** The tile set shown in Figure 2b forms only one type of assembly. A tile set consisting of the tiles in (b) and the four tiles shown here allows four types of assemblies to be formed. The vertical column of each type contains a different 2-bit binary sequence.

Under algorithmic assembly conditions [19], the assembly of DNA tiles into a crystal is only energetically favorable when it occurs cooperatively, i.e. by the formation of two or more sticky end bonds. The attachment of a tile to a crystal performs a step of a computation in the sense that a unique tile (among many possible in solution) may attach at a particular growth location. With an appropriate choice of tiles, DNA tile assembly can perform universal computation [18, 2].

The zig-zag crystal shown in Figure 2a is formed from the tiles shown in Figure 2b. Matching rules determine which tile fits where. When a zig-zag crystal is added to a solution of free tiles under algorithmic assembly conditions, growth is constrained to occur in a zig-zag pattern by the requirement that each tile addition must form two or more sticky end bonds, as shown in Figure 2a. It is easy to confirm that under such conditions, there is always a unique tile that may be added on each end of the ribbon.

Zig-zag crystals are designed so that under algorithmic assembly conditions, growth produces one new row at a time, and continued growth repeatedly copies a sequence. The requirement that a tile must attach by two bonds means that it must match both its vertical neighbor (another tile that is part of the new column being assembled), and its horizontal neighbor (in a previously assembled row). Several tiles might match the label on the vertical neighbor, but because tiles must make two correct bonds in order to join the assembly, only a tile that also matches the label on the horizontal neighbor can be added. Therefore, the tile being added in the new column must correspond to the one in previous column. As a result, information is inherited through templated growth. The set of tiles formed by adding the tiles in Figure 2c to those shown in Figure 2b can propagate one of four strings. Additional tiles may be added to the set of tiles in Figures 2b and 2c to create a tile set that copies one of 2^n sequences of width n. We will later discuss tile sets in which an unbounded amount of information can be copied.

The growth of a zig-zag DNA crystal increases the number of copies of the original information present in the ribbon, but does not change the rate at which new copies of the sequence are produced. The rate of copying can be sped up by shear forces that cause crystals to break. With each new crystal that is created by breakage, two new sites become available to copy information. Repeated applications of shear force interspersed with time to grow therefore exponentially amplify an initial piece of information. Occasionally, a tile matching only one bond rather than two will join the assembly, resulting in occasional copying errors, which are also inherited. If errors happen during copying, which they will under almost any achievable condition [19], and crystals with particular sequences grow faster than others, then evolution can occur.

3 The Royal Road

A selection experiment for DNA crystal evolution involves both an environment (available resources and laws of chemistry and physics) and DNA crystals that grow and reproduce within that environment. Artificial evolution experiments must set up both. Here, a set of DNA tiles is used to define an environment for

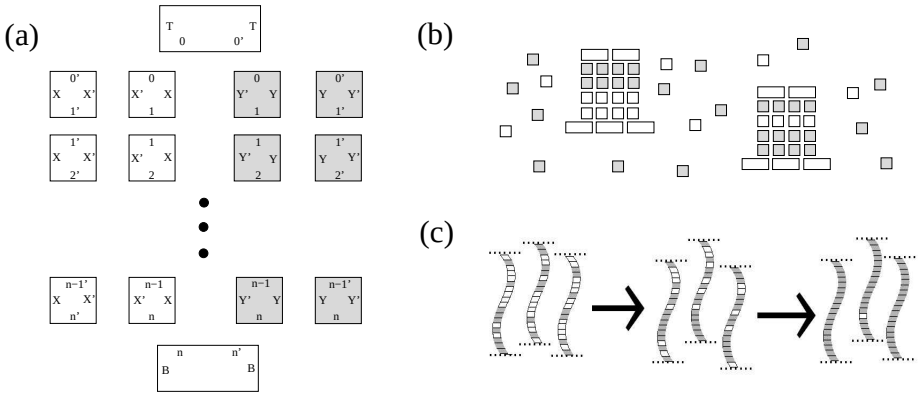

Fig. 3. The royal road tile set. (a) The royal road tile set consists of four tiles for each of n sequence positions, two for propagating an X bit and two for propagating a Y bit. Two boundary tiles are also used. 2^n different sequences can be copied with this tile set. **(b)** When more Y than X tiles are present, sequences containing more Y tiles tend to grow faster. **(c)** As growth progresses, sequences containing mostly Y tiles become more and more common. Each sequence shown represents an assembly consisting of many copies of the illustrated sequence.

crystal growth. The set of DNA tiles determines the set of sequences which may be copied and the "chemistry" of the system, i.e. the rules which tiles bind to each other [1]. A particular arrangement of DNA tiles is the information that is propagated in these experiments, the genotype; it is the organism being evolved. The phenotype of a sequence is its replication rate in the given environment. In this section we describe a tile set that allows many kinds of sequences to grow; a selection pressure results from physical conditions in which tile concentrations differ for each tile type.

A DNA crystal can grow only when it comes in contact with a tile that can be added favorably to the crystal. In a well-mixed reaction vessel, the higher the concentrations of tiles of the type that may be legally added, the more quickly such contact occurs. Therefore, a simple selection pressure results from a difference in concentration between tile types used to copy the sequence information: assemblies with sequences containing tiles present at high concentrations will grow and reproduce faster than assemblies with sequences containing tiles present at very low concentrations.

A tile set in which one of two bits can be propagated at each of n sequence positions is shown in Figure 3a. Let X_i and Y_i be the two tile types that can be propagated at position i. If Y_i is present in solution at a concentration higher than that of X_i, as in Figure 3b, the fitness landscape for this selection resembles

[1] Our choice of terminology reflects the observation that whether a self-replicator is made from clay, biological polymer or other material, the chemistry of the specific elements involved determines the evolutionary landscape. As an example, the chemistry of nucleic acids can make some sequences hard to copy. Certain sequences fold up or bulge [9], making copying of those sections more difficult. Here, the constraints are not on how a sequence folds, but on how its elements fit together: the tile set similarly determines the evolutionary landscape.

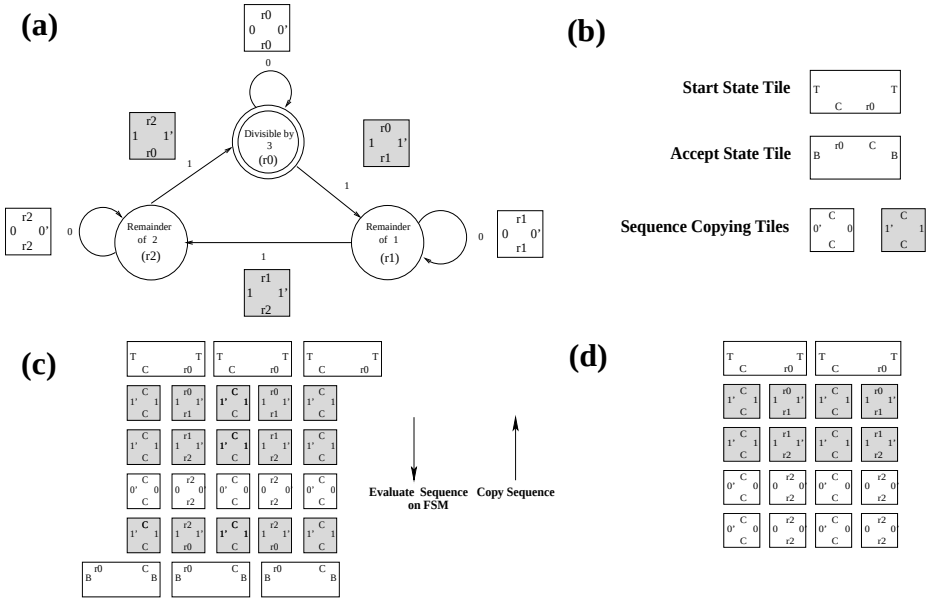

Fig. 4. Selection of sequences with particular numbers of logical 1's. (a) A diagram of a finite state machine that can determine whether a binary sequence contains a number of 1's that is divisible by 3. The double circled state is both the start and the accept state. (In general these states are not the same.) The tiles shown can be used with the tiles in (b) to follow the instructions of the machine during tile assembly. **(b)** Additional tiles needed to complete the tile set in (a). The construction shown in (a) and (b) can be generalized to any finite state machine. **(c)** An assembly encoding a sequence accepted by the machine in (a). Evaluation ends in an accept state, so a bottom tile may be added and assembly can continue. **(d)** An assembly encoding a sequence not accepted by the finite state machine in (a). Because execution of the finite state machine ended with a state other than the accept state, assembly cannot continue.

the simplest case of a well-studied problem in genetic algorithms, the "royal road" [12]. Here, the growth rate increases monotonically with the number of Y_i's in the sequence s. So long as the Y_i tiles remain more common in solution, sequences containing only Y_i tiles will quickly dominate (Figure 3c).

4 Selection of Regular Languages

Section 3 illustrated how tile concentration can create a selection pressure, causing some sequences to grow faster than others. While this is a simple selection pressure to understand, the adaptation that occurs is also simple. In this section we describe how a single tile set allows for the replication of an infinite number of sequences and how sequence constraints imposed by the tile set can provide more interesting selection pressures.

In the previous example, the "chemistry" of the tile set determined the length of sequences that could be copied, and which tiles could be used in which position of the sequence. Here we consider evolution when the tile set "chemistry"

allows only certain sequences to be copied, but they may be arbitrarily long. In particular, the tile set environment allows copying only of sequences that are accepted by a particular finite state machine.

A finite state machine is an abstract device that can perform a computation requiring only a fixed amount of memory. It consists of a set of states and rules describing how to transition between states as each character of input is received. Computation begins in a prescribed state. When the inputs have all been received, the current state is either in an accept state, in which case the the input is accepted, or a reject state. Figure 4a shows a simple finite state machine (along with the tiles that implement the transition steps of the machine) which detects whether the number of ones in a binary sequence input is divisible by three.

The self-assembly of DNA sequence [1] and tile [13] alphabets can generate the set of sequences accepted by a given finite state machine, also known as a regular language. Accepted sequences can be of any length. In contrast to the tile sets described in Section 3, where the top and bottom sides of a tile encode the position in the fixed-length sequence where the tile can be added, the top and bottom sides of the tiles in Figure 4a encode the state of the machine as it processes each character of the sequence being copied.

A tile set that copies only inputs accepted by a given finite state machine is constructed as follows. Each possible transition between states is encoded as a single tile (Figure 4a). The left and right sides of the tile encode the input, the top side encodes the state that machine is in before the input is received and the bottom side encodes the state that the machine transitions to after the input has been received. The top boundary tile encodes the start state and a bottom boundary tile encodes each accept state (Figure 4b). Another set of tiles copies a sequence that has been accepted by the machine. These tiles have only one state on their bottom and top sides, and encode the same sequence bit on their left and right sides.

During growth down the crystal[2], assembly evaluates the sequence according to the finite state machine's rules. If the machine ends in an accept state, a bottom tile can bind to the site and upward growth can begin (Figure 4c). If the machine is not in an accept state, no bottom tile exists which matches the growth front, and growth stops (Figure 4d). Thus, only sequences which are accepted by the machine will continue to be replicated. These sequences will be the ones that are selected for.

More complex selection pressure results if the crystals grown in this tile set environment are moved to an environment containing tiles that accept a different language of sequences. For example, crystals grown using the tiles shown here

[2] Growth on the left side of the zig-zag crystal in Figure 4c reads the sequence elements backward, and evaluates the finite state machine in reverse. While running the finite state machine shown in Figure 4a backward accepts the same set of states as running the machine forward, for other machines there may be non-determinism when the machine is run in reverse. A step may be possible that cannot lead to the start state, leaving an uncompleted assembly. Assemblies corresponding to tile sets of this type will grow mostly in the direction where the finite state machine is evaluated in the correct direction. With some additional complexity, it is also possible to replace this tile set an equivalent tile set that can grow only in the forward direction [17].

might be moved to a mixture containing tiles that allowed only sequences with a number of ones that is divisible by 5 to grow. Only sequences with a number of ones divisible by 15 could survive in both environments.

5 Acceptable Error Rates for DNA Tile-Based Evolution

While several experimental studies have shown that DNA tiles can process information through cooperative binding [11, 15], it is also becoming clear that errors occur often during algorithmic assembly [15]. This is a concern because a low error rate is vital to the design of a self-replicator. If the error rate exceeds an error threshold [7], genetic meltdown occurs and sequences become totally random. In this section we describe how to decrease the error rate below any relevant error threshold.

Errors during assembly occur when a tile binds to a growing assembly by fewer than two bonds, an event called an unfavorable attachment. A mismatch error, an unfavorable attachment that only partially matches the adjacent tiles, causes an error in replication (Figure 5a). Additionally, in the absence of a pre-existing crystal, a series of unfavorable attachments occasionally produces a full-width crystal with a random sequence, an event called spontaneous nucleation.

Both these kinds of errors can be analyzed using a reversible model of DNA tile self-assembly based on the physics and chemistry of DNA hybridization [19]. Prior work on the robustness of algorithmic self-assembly in this model can be adapted in order to show that, at a moderate cost of tile set complexity and assembly speed, mismatch error rates can be made as small as is desired. "Proofreading" tile sets implement the same logic of an original tile set but assemble more robustly, dramatically reducing mismatch error rates without significant slow-down [20, 6, 14]. The general idea of proofreading is to redundantly encode each element of sequence. When the proofreading method is applied to the zig-zag tile set (Figure 5b), correct tile additions are stabilized by additional tiles in the same block that encode the same sequence element, whereas several incorrect additions instead of just one are needed to propagate a sequence element incorrectly (Figure 5c). Error rates decrease exponentially as larger blocks of proofreading tiles are used [20].

Similar error correction techniques also exist for the prevention of spontaneous nucleation errors. Like other crystallization processes, the rate at which spontaneous nucleation of growing zig-zag assemblies occurs is dependent on the energy of the critical nucleus for growth. For zig-zag crystals, this critical nucleus is a small assembly that contains both a top and bottom boundary tile. By increasing the minimum width of an assembly that can contain both these tiles, it is possible to increase the energy of the critical nucleus. For example, the rate of spontaneous nucleation of the zig-zag tile set shown in Figure 3a decreases exponentially with the width n [16]. We expect that the same qualitative result applies to the more complex tile sets described in this paper.

6 Conclusions

To study the physical principles of Darwinian evolution, we propose a physical system based on DNA crystals in which a combinatorial variety of genotypes can

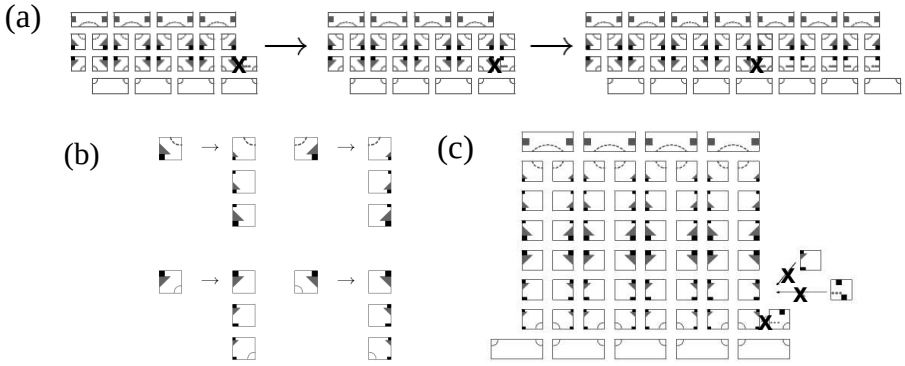

Fig. 5. Proofreading for zig-zag assembly. (a) Kinetic trapping is the major cause of mismatch errors in DNA tile assembly. When a tile attaches to an assembly on only one side, it forms a low energy bond and usually dissociates quickly. However, if another tile attaches to the assembly at an adjacent location before it can dissociate, the tile may be trapped. The mutated sequence will be copied to subsequent columns. **(b)** Zig-zag proofreading transformations of the four zig-zag middle tiles in Figures 2b. **(c)** Zig-zag assembly of the original sequences using the transformed tile set. When a single tile that produces an error attaches to the assembly, either the tile must fall off and be replaced by the correct tile, or further errors are necessary in order to continue growth.

be faithfully replicated and a genotype can direct a behavior or other measurable parameter that can be subject to selection. DNA crystals are simple, containing no biological parts, and can be programmed to replicate an infinite variety of genotypes. The ability to program the interactions between tiles allows us to induce selection pressures which favor the growth of assemblies with interesting properties. Error correction techniques exist which can lower the replication error rate as much as is required to avoid genetic meltdown, at the cost of a small amount of additional complexity.

Acknowledgements: We would like to thank Bernie Yurke, Gerald Joyce, Andy Ellington, Graham Cairns-Smith, Paul Rothemund, Dave Zhang and David Soloveichik for helpful discussion on sequence amplification and evolution. The AFM image in the inset of Figure 1c was taken by Ho-Lin Chen. This research was partially supported by NSF awards #0093486 and #0432193.

References

1. Leonard M. Adleman. Molecular computation of solutions to combinatorial problems. *Science*, 266:1021–1024, November 11, 1994.
2. Leonard M. Adleman, Jarkko Kari, Lila Kari, and Dustin Reishus. On the decidability of self-assembly of infinite ribbons. *Symposium on Foundations of Computer Science (FOCS)*, 43:530, 2002.
3. A. Graham Cairns-Smith. The origin of life and the nature of the primitive gene. *Journal of Theoretical Biology*, 10:53–88, 1966.

4. A. Graham Cairns-Smith. The chemistry of materials for artificial Darwinian systems. *International Revs. Phys. Chem.*, 7:209–250, 1988.
5. Nickolas Chelyapov, Yuriy Brun, Manoj Gopalkrishnan, Dustin Reishus, Bilal Shaw, and Leonard Adleman. DNA triangles and self-assembled hexagonal tilings. *J. Am. Chem. Soc.*, 126(43):13924–13925, 2004.
6. Ho-Lin Chen and Ashish Goel. Error free self-assembly using error prone tiles. In *DNA Computing 10*, Berlin Heidelberg, 2004. Springer-Verlag.
7. Manfred Eigen. Self-organization of matter and evolution of biological macromolecules. *Naturwissenschaften*, 58(10):465–523, 1971.
8. Tsu-Ju Fu and Nadrian C. Seeman. DNA double-crossover molecules. *Biochemistry*, 32:3211–3220, 1993.
9. Gerald F. Joyce. Nonenzymatic template-directed synthesis of informational macromolecules. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 52, pages 41–51, 1987.
10. Thomas H. LaBean, Hao Yan, Jens Kopatsch, Furong Liu, Erik Winfree, John H. Reif, and Nadrian C. Seeman. Construction, analysis, ligation and self-assembly of DNA triple crossover complexes. *J. Am. Chem. Soc.*, 122(9):1848–1860, 2000.
11. Chengde Mao, Thomas H. LaBean, John H. Reif, and Nadrian C. Seeman. Logical computation using algorithmic self-assembly of DNA triple-crossover molecules. *Nature*, 407(6803):493–496, 2000.
12. Melanie Mitchell, Stephanie Forrest, and John H. Holland. The royal road for genetic algorithms: Fitness landscapes and GA performance. In *Proceedings of the First European Conference on Artificial Life*, 1992.
13. John H. Reif. Local parallel biomolecular computation. In Harvey Rubin and David Harlan Wood, editors, *DNA Based Computers III*, volume 48 of *DIMACS*, pages 217–254, Providence, RI, 1997. American Mathematical Society.
14. John H. Reif, Sudheer Sahu, and Peng Yin. Compact error-resilient computational DNA tiling assemblies. In *DNA Computing 10*, Berlin Heidelberg, 2004. Springer-Verlag.
15. Paul W. K. Rothemund, Nick Papadakis, and Erik Winfree. Algorithmic self-assembly of DNA Sierpinski triangles. *PLOS Biology*, 2:424–436, 2004.
16. Rebecca Schulman and Erik Winfree. Controlling nucleation rates in algorithmic self-assembly. In *DNA Computing 10*, Berlin Heidelberg, 2004. Springer-Verlag.
17. Erik Winfree. Self healing tile sets for algorithmic self-assembly. In preparation.
18. Erik Winfree. On the computational power of DNA annealing and ligation. In Richard J. Lipton and Eric B. Baum, editors, *DNA Based Computers*, volume 27 of *DIMACS*, pages 199–221, Providence, RI, 1996. American Mathematical Society.
19. Erik Winfree. Simulations of computing by self-assembly. Technical Report CS-TR:1998.22, Caltech, 1998.
20. Erik Winfree and Renat Bekbolatov. Proofreading tile sets: Error-correction for algorithmic self-assembly. In Junghuei Chen and John Reif, editors, *DNA Computing 9*, volume LNCS 2943, pages 126–144, Berlin Heidelberg, 2004. Springer-Verlag.
21. Erik Winfree, Furong Liu, Lisa A. Wenzler, and Nadrian C. Seeman. Design and self-assembly of two-dimensional DNA crystals. *Nature*, 394:539–544, 1998.